

Cassette Search User Guide

The Cassette Search tool enables the user to find conserved gene neighborhoods across multiple isolate genomes. User defines the desired functions or “hooks” in the input page along with other parameters such as distance between hooks. There are two ways to access the tool from the IMG user interface (UI):

- From Workspace->My Data Sets->Genome:
- From Find Genes->Cassette Search:

Cassette Search NEW

Cassette search looks for **conserved gene neighborhoods** in selected genomes using a set of selected functions as “hooks”.
Required functions may be grouped in parentheses to specify that these must be on the same gene. e.g. (COG0232,pfam01966,pfam13286).

Search: ☐ All public isolate genomes ☒ Only genomes in selected workspace dataset:

| | | |
|--|----------------------|--|
| Required Hooks (required, min 3, max 10): | <input type="text"/> | e.g. (COG0232,pfam01966,pfam13286),KO:K06147,TIGR00861 |
| Additional Hooks (optional): | <input type="text"/> | e.g. pfam00535,pfam00664,KO:K02025,EC:3.2.1.156 |
| Minimum Number of Additional Hooks (optional): | <input type="text"/> | Number, min: 0, max: 10 |
| Maximum Distance between Hooks: | <input type="text"/> | Nucleotides (default: 5000, max: 20000) |
| Extend Boundaries by: | <input type="text"/> | Nucleotides (default: 5000, max: 20000) |
| Minimum Distance from Scaffold Edge: | <input type="text"/> | Nucleotides (min: 1) |
| Name Your Search (required) | <input type="text"/> | Up to 50 characters, no spaces or special characters |
| Comment | <input type="text"/> | Additional comment |

ResetSearch

hint: The number of results returned will be limited to 1000 rows.
Use the following illustration as a guideline to this tool.

The diagram illustrates the cassette search process. It shows a sequence of genomic elements: a dashed line for 'Boundary Extension', followed by an orange box for 'Additional (optional)', a blue box for 'Required', another blue box for 'Required', a third blue box for 'Required', an orange box for 'Additional (optional)', and another dashed line for 'Boundary Extension'. A bracket labeled 'Max distance' spans the three 'Required' boxes.

The new Cassette Search function is an extension of the ClusterScout tool originally developed for IMG/ABC (<https://academic.oup.com/nar/article/45/D1/D560/2605822>).

The Cassette Search function allows users to search through all public isolate genomes in the IMG database, or a user-curated list of genomes (**isolates only**) stored in a workspace “genome set”. You can learn how to add genomes as a “Genome Set” in Workspace [here](#).

Please note that the cassette search results will be **limited to 1000 cassettes** at most.

The search parameters are as follows:

- **Required Hooks:** Users should specify 3 to 10 functions (comma separated) in the search. The functions can be:
 - COG functions such as: COG0396

- Pfam functions such as: pfam00535
- TIGRFam functions such as: TIGR00861
- Enzymes such as: EC:3.2.1.156
- KO terms such as: KO:K02025

Moreover, users can specify functions within the same genes by using parentheses. For example, (KO:K14331,pfam11266) means that a single gene must be annotated with both functions within parentheses. *This is particularly useful when multi-domain proteins are being used. Or if there are potential differences in sensitivity of functional annotation types.*

- Additional Hooks (optional): Any of the additional COG, Pfam, TIGRfam, Enzymes and/or KO terms to be included in the cassette.
- Minimum Number of Additional Hooks (optional): The cassette must include at least this number of additional hooks.
- Maximum Distance between Hooks: maximal number of bases allowed between two required hooks
- Extend Boundaries by: The number of bases to be added to both ends of the cassette to extend the region retrieved around the hooks.
- Minimum Distance from Scaffold Edge: The number of bases from the cassettes to both edges of the scaffold (possibly desirable when working with highly fragmented draft genomes)
- Name Your Search (required): The job name for this search. After the computation is done, users will receive an email alert. OR Users can go to **MyJob** to find a particular cassette search result by name.
- Comment (optional): Optional free text comment. (keep notes for future reference)

Please note that the search results will be **limited to at most 1000 cassettes**.

Example 1: Find *glycocin* cassettes in *Bacillus*

Search Pfams: pfam00005, pfam00535, pfam03412, pfam00664 (representing ABC transporter, Glycosyl transferase family 2, Peptidase C39 family and ABC transporter transmembrane region, respectively) in all public Bacillus genomes in IMG.

First, create a workspace genome dataset to include all public Bacillus genomes in IMG. Let's name the genome dataset "Bacillus."

Then run cassette/conserved neighborhood search using the above mentioned functions and dataset, and use the default parameters:

Genome Set(s): Bacillus

Required Hooks: pfam00005,pfam00535,pfam03412,pfam00664

Additional Hooks: At least 4 of (pfam00664,pfam00535,pfam00005,pfam03412)

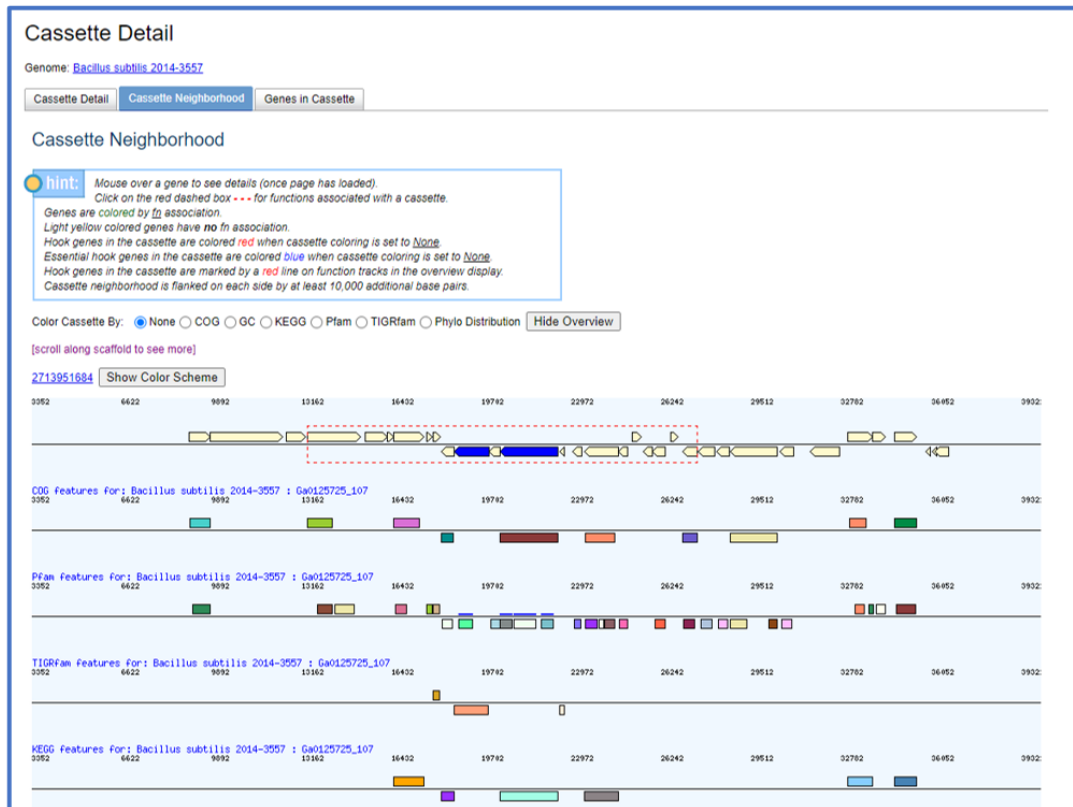
Minimum Number of Additional Hooks: 4

Maximum Distance between Hooks: 5000 nt

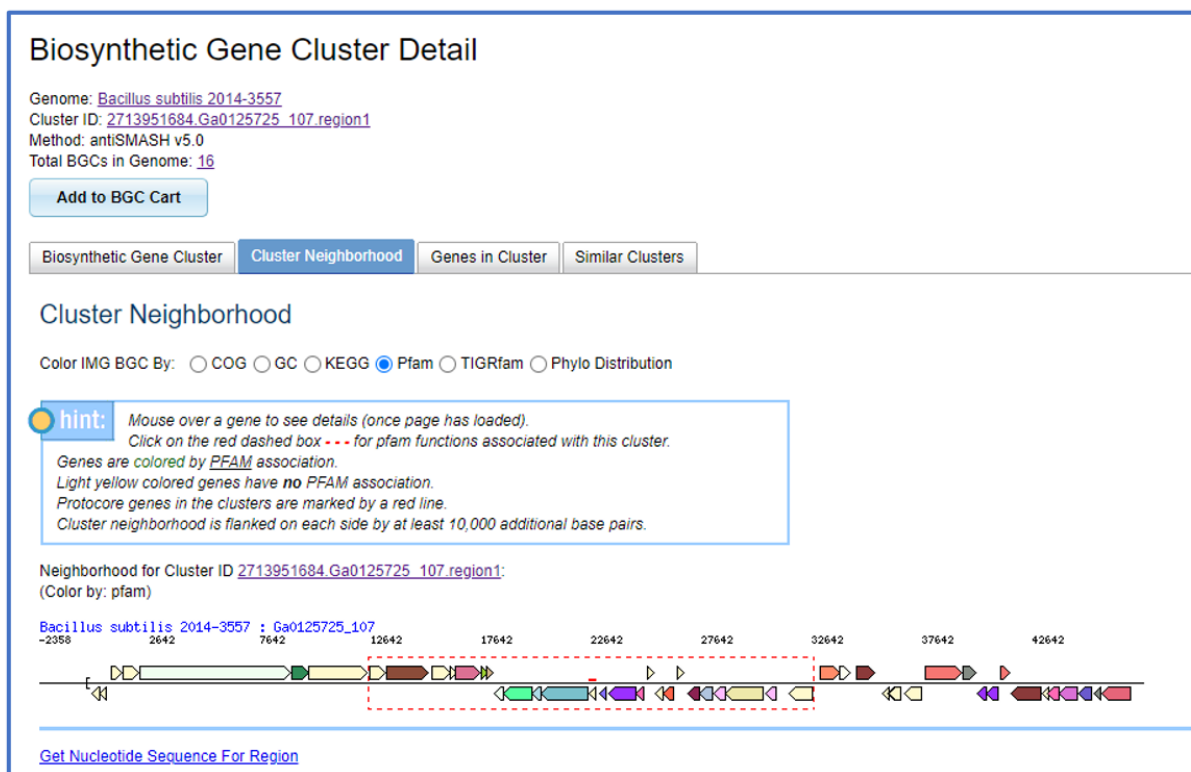
Extend Boundaries by: 5000 nt

Minimum distance from scaffold edge: 1000
Comment: user guide example 1 Bacillus genomes only

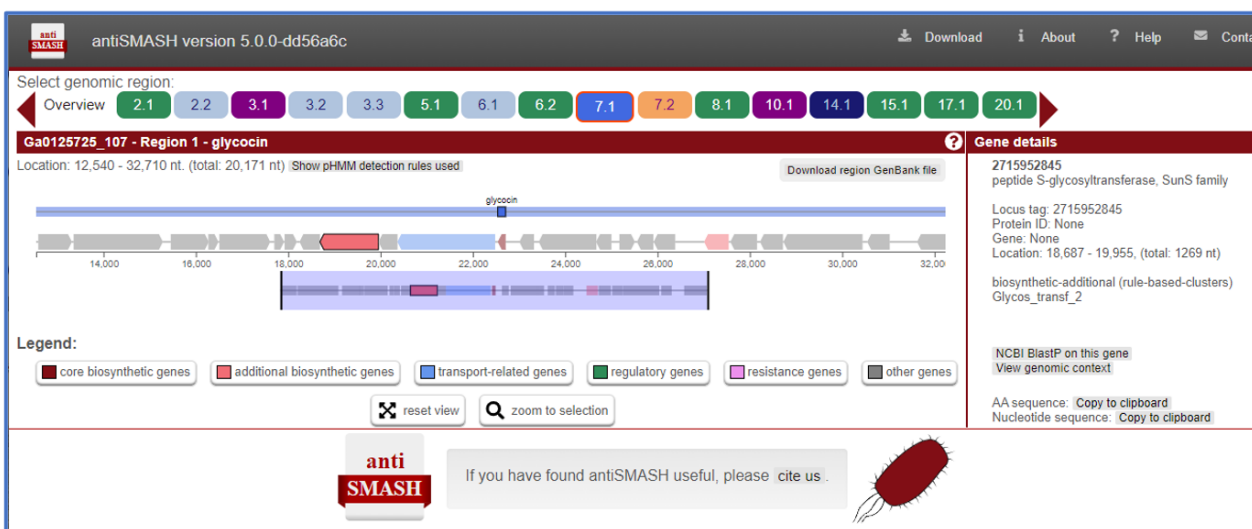
The search returns 87 results. Example of a cassette found in *Bacillus subtilis* 2014-3557 is shown below:



It corresponds to a *glycocin* biosynthetic gene cluster 2713951684.Ga0125725_107.region1 in IMG/ABC:



The corresponding antiSMASH v.5 prediction is as follows:



Example 2: Find alkane biosynthesis “operon” in *Cyanobacteria*

Assume we are interested in finding a two gene “operon” like fatty aldehyde decarbonylase (Ado) and acyl-ACP reductase (Aar) responsible for alkane synthesis in *Cyanobacteria*.

The first step will be to search and save all phylum Cyanobacteria genomes in IMG into a Workspace genome set.

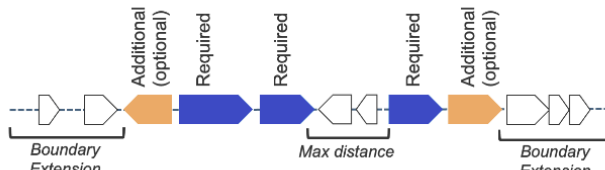
Then we can use TWO annotations available for Ado (**KO:K14331** fatty aldehyde decarboxylase AND **pfam11266** - Ald_deCOase) and one for Aar - **KO:K14330** (fatty aldehyde-generating acyl-ACP reductase) as our minimum THREE required hooks.

Cassette Search NEW

Cassette search looks for conserved gene neighborhoods in selected genomes using a set of selected functions as "hooks".
Required functions may be grouped in parentheses to specify that these must be on the same gene. e.g. (COG0232, pfam01966, pfam13286).
Search: ☐ All public isolate genomes ☒ Only genomes in selected workspace dataset: Cyanobacteria

| | | |
|--|-----------------------------------|--|
| Required Hooks (required, min 3, max 10): | (KO:K14331, pfam11266), KO:K14330 | e.g. (COG0232, pfam01966, pfam13286), KO:K06147, TIGR00861 |
| Additional Hooks (optional): | | e.g. pfam00535, pfam00664, KO:K02025, EC:3.2.1.156 |
| Minimum Number of Additional Hooks (optional): | 0 | Number, min: 0, max: 10 |
| Maximum Distance between Hooks: | 5000 | Nucleotides (default: 5000, max: 20000) |
| Extend Boundaries by: | 5000 | Nucleotides (default: 5000, max: 20000) |
| Minimum Distance from Scaffold Edge: | 1000 | Nucleotides (min: 1) |
| Name Your Search (required): | example2 | Up to 50 characters, no spaces or special characters |
| Comment | user guide example 2 | Additional comment |

hint: The number of results returned will be limited to 1000 rows.
Use the following illustration as a guideline to this tool.



We will only search genomes in the Cyanobacteria workspace dataset, and the result will be saved to a job called "example2." When the computation is done, IMG will send an email notification linking to the result:

Cassette Search Results

Job Name: example2

Genome Set(s): Cyanobacteria

Required Hooks: (KO:K14331.pfam11266),KO:K14330

Additional Hooks: At least 0 of ()

Maximum Distance between Hooks: 5000 nt

Extend Boundaries by: 5000 nt

Minimum Distance from Scaffold Edge: 1000

Comment: user guide example 2

Showing 1 to 10 of 867 entries

| | Cassette | Gene Count | Genome Name | Scaffold | Start Coord | End Coord | Length |
|--------------------------|--|--|---|--|---|---|--|
| | <input type="text" value="Search Cassette"/> | <input type="text" value="Search Gene Count"/> | <input type="text" value="Search Genome Name"/> | <input type="text" value="Search Scaffold"/> | <input type="text" value="Search Start Coord"/> | <input type="text" value="Search End Coord"/> | <input type="text" value="Search Length"/> |
| <input type="checkbox"/> | 1 | 19 | Prochlorococcus sp. AG-363-A16 / (Screened) | 2667535442 | 9236 | 27625 | 18390 |
| <input type="checkbox"/> | 2 | 14 | Gloeocapsa sp. PCC 73106 | 2508503451 | 46690 | 63927 | 17238 |
| <input type="checkbox"/> | 3 | 10 | Pseudanabaena sp. FACHB-1277 | 2909891102 | 38597 | 55540 | 16944 |
| <input type="checkbox"/> | 4 | 16 | Limnospira indica PCC 8005 | 2751243480 | 2803660 | 2820315 | 16656 |
| <input type="checkbox"/> | 5 | 21 | Limnospira fusiformis SAG 85.79 | 2883428434 | 2817868 | 2834523 | 16656 |
| <input type="checkbox"/> | 6 | 15 | Limnospira indica PCC 8005 | 646280014 | 112700 | 129354 | 16655 |
| <input type="checkbox"/> | 7 | 21 | Arthrospira platensis C1 | 2507281629 | 5361416 | 5378064 | 16649 |
| <input type="checkbox"/> | 8 | 13 | Iningainema sp. BLCCT55 | 2914081501 | 22595 | 39116 | 16522 |
| <input type="checkbox"/> | 9 | 11 | Calothrix sp. 336/3 | 2654629099 | 5377911 | 5393905 | 15995 |
| <input type="checkbox"/> | 10 | 12 | Fischerella thermalis WC441 | 2802497020 | 31366 | 47287 | 15922 |

Showing 1 to 10 of 867 entries

Save Cassettes to My Workspace

hint: Even though you can save large amount of data into workspace, many profile functions will timeout for extremely large workspace datasets

Save selected Cassettes to [My Workspace](#).

(Special characters in file name will be removed and spaces converted to _)

Click “Select All”, then “Save” to save all the cassettes from this job as a Cassette Set for further analysis. These sets can later be found under the “Workspace” menu.

My Workspace - Individual Cassette Set

Set Name: Cyanobacteria

Cassettes in Set Neighborhoods Function Profile

Add Scaffolds of Selected to Cart

Add Genes of Selected to Cart

Showing 1 to 10 of 867 entries 3 rows selected

| | Cassette | Genome Name | Scaffold | Start Coord | End Coord | Length | Required Hooks | Ad |
|-------------------------------------|--|---|--|---|---|--|--|--|
| | <input type="text" value="Search Cassette"/> | <input type="text" value="Search Genome Name"/> | <input type="text" value="Search Scaffold"/> | <input type="text" value="Search Start Coord"/> | <input type="text" value="Search End Coord"/> | <input type="text" value="Search Length"/> | <input type="text" value="Search Required Hooks"/> | <input type="text" value="Search Ad"/> |
| <input type="checkbox"/> | 1 | Prochlorococcus sp. AG-363-A16 / (Screened) | 2667535442 | 9236 | 27625 | 18390 | (KO:K14331.pfam11266),KO:K14330 | |
| <input type="checkbox"/> | 2 | Gloeocapsa sp. PCC 73106 | 2508503451 | 46690 | 63927 | 17238 | (KO:K14331.pfam11266),KO:K14330 | |
| <input type="checkbox"/> | 3 | Pseudanabaena sp. FACHB-1277 | 2909891102 | 38597 | 55540 | 16944 | (KO:K14331.pfam11266),KO:K14330 | |
| <input checked="" type="checkbox"/> | 4 | Limnospira indica PCC 8005 | 2751243480 | 2803660 | 2820315 | 16656 | (KO:K14331.pfam11266),KO:K14330 | |
| <input checked="" type="checkbox"/> | 5 | Limnospira fusiformis SAG 85.79 | 2883428434 | 2817868 | 2834523 | 16656 | (KO:K14331.pfam11266),KO:K14330 | |
| <input checked="" type="checkbox"/> | 6 | Limnospira indica PCC 8005 | 646280014 | 112700 | 129354 | 16655 | (KO:K14331.pfam11266),KO:K14330 | |
| <input type="checkbox"/> | 7 | Arthrospira platensis C1 | 2507281629 | 5361416 | 5378064 | 16649 | (KO:K14331.pfam11266),KO:K14330 | |
| <input type="checkbox"/> | 8 | Iningainema sp. BLCCT55 | 2914081501 | 22595 | 39116 | 16522 | (KO:K14331.pfam11266),KO:K14330 | |
| <input type="checkbox"/> | 9 | Calothrix sp. 336/3 | 2654629099 | 5377911 | 5393905 | 15995 | (KO:K14331.pfam11266),KO:K14330 | |
| <input type="checkbox"/> | 10 | Fischerella thermalis WC441 | 2802497020 | 31366 | 47287 | 15922 | (KO:K14331.pfam11266),KO:K14330 | |

Showing 1 to 10 of 867 entries 3 rows selected

Add Scaffolds of Selected to Cart

Add Genes of Selected to Cart

Save Cassettes to My Workspace

hint: Even though you can save large amount of data into workspace, many profile functions will timeout for extremely large workspace datasets

For example, you can view neighborhoods or function profile of selected cassettes in a cassette set. You can also save scaffolds or genes of selected cassettes.

Neighborhoods for Selected Cassettes

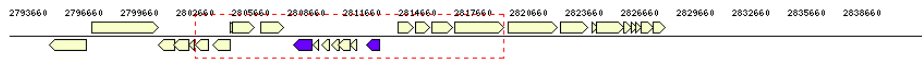
Cassette Set Name: [Cyanobacteria](#)

hint: Mouse over a gene to see details (once page has loaded).
Hook **genes** in the cassette are colored **blue** (required hook), **red** (additional hook), or **purple** (both hooks found on same gene) when cassette coloring is set to *None*.
Functions used as "hooks" are marked by **blue** (required hook) or **red** (additional hook) line on function tracks in the overview display.
Cassette neighborhood is flanked on each side by at least 10,000 additional base pairs.
Note: When selected "hooks" fall on the same gene, it may appear as though only a single hook is found. In that case, looking at the function tracks is more useful.

Color Cassette By: ☒ None ☐ COG ☐ GC ☐ KEGG ☐ Pfam ☐ TIGRfam ☐ Phylo Distribution [Show Scaffold Overview](#)

[2751243490](#)

[Show Color Scheme](#)



Color Cassette By: ☒ None ☐ COG ☐ GC ☐ KEGG ☐ Pfam ☐ TIGRfam ☐ Phylo Distribution [Show Scaffold Overview](#)

[2883428434](#)

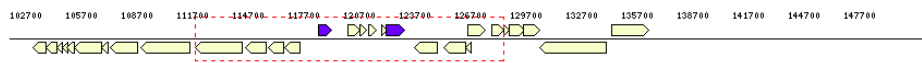
[Show Color Scheme](#)



Color Cassette By: ☒ None ☐ COG ☐ GC ☐ KEGG ☐ Pfam ☐ TIGRfam ☐ Phylo Distribution [Show Scaffold Overview](#)

[648280014](#)

[Show Color Scheme](#)



Function Profile

Showing genes with **Pfam** for selected cassette(s) in set [Cyanobacteria](#)

hint: Please note: cassette names in headers are assigned for display only

Showing 1 to 10 of 16 entries

First Previous [1](#) [2](#) Next Last [Export](#) [Select All](#) [Clear All](#) [Select - page](#) [Deselect - page](#) [Column Selector](#) Show [10](#)

| Function ID | Function Name | Total selected in: Cyanobacteria | C1 | C2 | C3 |
|--|---|----------------------------------|-------------------|-------------------|-------------------|
| <input type="checkbox"/> pfam00211 | Adenylate and Guanylate cyclase catalytic domain | 3 | 1 | 1 | 1 |
| <input type="checkbox"/> pfam00239 | Resolvase, N terminal domain | 3 | 1 | 1 | 1 |
| <input type="checkbox"/> pfam00376 | MerR family regulatory protein | 3 | 1 | 1 | 1 |
| <input type="checkbox"/> pfam00528 | Binding-protein-dependent transport system inner membrane component | 3 | 1 | 1 | 1 |
| <input type="checkbox"/> pfam00534 | Glycosyl transferases group 1 | 3 | 1 | 1 | 1 |
| <input type="checkbox"/> pfam00535 | Glycosyl transferase family 2 | 3 | 1 | 1 | 1 |
| <input type="checkbox"/> pfam00764 | Argininosuccinate synthase | 3 | 1 | 1 | 1 |
| <input type="checkbox"/> pfam00989 | PAS fold | 3 | 1 | 1 | 1 |
| <input type="checkbox"/> pfam01590 | GAF domain | 3 | 1 | 1 | 1 |
| <input type="checkbox"/> pfam01844 | HNH endonuclease | 3 | 1 | 1 | 1 |

First Previous [1](#) [2](#) Next Last [Export](#) [Select All](#) [Clear All](#) [Select - page](#) [Deselect - page](#) [Column Selector](#)

Showing 1 to 10 of 16 entries

[Add Selected to Function Cart](#) [Add Selected to Gene Cart](#)

From your cassette search job or from a saved cassette set, you can click on an individual Cassette ID to see the detailed information for that cassette.

In the Cassette Detail page, there is a Cassette Neighborhood tab to show the cassette with genes. Click on "Show Scaffold Overview" to see gene function annotations with all required hooks underlined:

Cassette Detail

Genome: [Gloeocapsa sp. PCC 73106](#)

[Cassette Detail](#) [Cassette Neighborhood](#) [Genes in Cassette](#)

Cassette Neighborhood

hint: Mouse over a gene to see details (once page has loaded).
Hook genes in the cassette are colored **blue** (required hook), **red** (additional hook), or **purple** (both hooks found on same gene) when cassette coloring is set to **None**.
Functions used as "hooks" are marked by **blue** (required hook) or **red** (additional hook) **line** on function tracks in the overview display.
Cassette neighborhood is flanked on each side by at least 10,000 additional base pairs.
Note: When selected "hooks" fall on the same gene, it may appear as though only a single hook is found. In that case, looking at the function tracks is more useful.

Color Cassette By: ☒ None ☐ COG ☐ GC ☐ KEGG ☐ Pfam ☐ TIGRFam ☐ Phylo Distribution [Hide Overview](#)

[\[scroll along scaffold to see more\]](#)

[2508503451](#) [Show Color Scheme](#)

